Multivariate location and dispersion
1 Introduction

Multivariate analysis deals with situations in which several variables are measured on each experimental unit.

In most cases of interest it is known or assumed that some form of relationship exists among the variables, and hence that considering each of them separately would entail a loss of information.

Some possible goals of the analysis are: reduction of dimensionality (principal components, factor analysis, canonical correlation); identification (discriminant analysis); explanatory models (multivariate linear model).

A $p$-variate observation is now a vector $x = (x_1, \ldots, x_p)' \in \mathbb{R}^p$ and a distribution $F$ means now a distribution on $\mathbb{R}^p$. 
In the classical approach, location of a $p$-variate random variable $\mathbf{x}$ is described by the expectation

$$\mu = \mathbb{E}\mathbf{x} = (\mathbb{E}x_1, \ldots, \mathbb{E}x_n)'$$

and dispersion is described by the covariance matrix

$$\text{Var}(\mathbf{x}) = \mathbb{E}\left((\mathbf{x} - \mu)(\mathbf{x} - \mu)'ight).$$

It is well known that $\text{Var}(\mathbf{x})$ is symmetric and positive semidefinite, and that for each constant vector $\mathbf{a}$ and matrix $\mathbf{A}$

$$\mathbb{E}(\mathbf{A}\mathbf{x} + \mathbf{a}) = \mathbf{A} \mathbb{E}\mathbf{x} + \mathbf{a}, \quad \text{Var}(\mathbf{A}\mathbf{x} + \mathbf{a}) = \mathbf{A}\text{Var}(\mathbf{x})\mathbf{A}'. $$
Classical multivariate methods of estimation are based on the assumption of an i.i.d. sample of observations \( X = \{x_1, \ldots, x_n\} \) with each \( x_i \) having a \( p \)-variate normal \( N_p(\mu, \Sigma) \) distribution with density

\[
f(x) = \frac{1}{(2\pi)^{p/2}\sqrt{|\Sigma|}} \exp \left( -\frac{1}{2}(x - \mu)' \Sigma^{-1} (x - \mu) \right),
\]

where \( \Sigma = \text{Var}(x) \) and \( |\Sigma| \) stands for the determinant of \( \Sigma \).

The contours of constant density are the elliptical surfaces

\[
\{ z : (z - \mu)' \Sigma^{-1} (z - \mu) = c \}.
\]

Assuming \( x \) multivariate normal implies that for any constant vector \( a \), all linear combinations \( a'x \) are normally distributed.
It also implies that since the conditional expectation of one coordinate with respect to any group of coordinates is a linear function of the latter, the type of dependence among variables is *linear*.

Thus methods based on multivariate normality will yield information only about linear relationships among coordinates.

As in the univariate case, the main reason for assuming normality is *simplicity*.

It is known that under the normal distribution, the MLE’s of $\mu$ and $\Sigma$ for a sample $x$ are respectively the sample mean and sample covariance matrix

$$
\bar{x} = \text{ave}(X) = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad \text{Var}(X) = \text{ave}\{(X - \bar{x})(X - \bar{x})'\}
$$
The sample mean and sample covariance matrix share the behavior of the distribution mean and covariance matrix under affine transformations, namely for each vector $a$ and matrix $A$

$$\text{ave}(AX + a) = A \text{ave}(X) + a, \quad \text{Var}(AX + a) = A \text{Var}(X) A',$$

where $AX + a$ is the dataset $\{Ax_i + a, i = 1, ..., n\}$.

This property is known as the \textit{affine equivariance} of the sample mean and covariances.

Just as in the univariate case, a few atypical observations may completely alter the sample means and/or covariances.

Worse still, a multivariate outlier need not be an outlier in any of the coordinates considered separately.
Example 1. The data from (Seber, 1984, Table 9.12) contains biochemical measurements on 12 men with similar weights. The data for phosphate and chloride are plotted in the next Figure.
Biochemical data

We see in the Figure that observation 3, which has the lowest phosphate value, stands out clearly from the rest.
However the next Figure, which shows the normal Q-Q plot of phosphate, does not reveal any atypical value, and the same occurs in the Q-Q plot of chloride (not shown).
Thus the atypical character of observation 3 is visible only when considering both variables simultaneously.
The Table below shows that omitting this observation has no important effect on means or variances, but the correlation almost doubles in magnitude.

That is, the influence of the outlier has been to decrease the correlation by a factor of two relative to that without the outlier.

<table>
<thead>
<tr>
<th></th>
<th>means</th>
<th>var.</th>
<th>correl.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>phos.</td>
<td>chlor.</td>
<td>phos.</td>
</tr>
<tr>
<td>Complete data</td>
<td>1.79</td>
<td>6.01</td>
<td>0.26</td>
</tr>
<tr>
<td>Without obs. 3</td>
<td>1.87</td>
<td>6.16</td>
<td>0.20</td>
</tr>
</tbody>
</table>

Here we have an example of an observation which is not a one-dimensional outlier in either coordinate but strongly affects the results of the analysis.

This example shows the need for robust substitutes of the mean vector and covariance matrix, which will be the main theme of this Lecture.
Let \((\hat{\mu}(X), \hat{\Sigma}(X))\) be location and dispersion estimates corresponding to a sample \(X = \{x_1, ..., x_n\}\). Then the estimates are \textit{affine equivariant} if

\[
\hat{\mu}(AX + b) = A\hat{\mu}(X) + b, \quad \hat{\Sigma}(AX + a) = A\hat{\Sigma}A'.
\]

Affine equivariance is a desirable property of an estimate.

This is however not a mandatory property, and may in some cases be sacrificed for other properties as e.g. computational speed.
1.1 Outlier detection

As in the univariate case, one may consider the approach of outlier detection.

The squared *Mahalanobis distance* between the vectors $\mathbf{x}$ and $\mu$ with respect to the matrix $\Sigma$ is defined as

$$d(\mathbf{x}, \mu, \Sigma) = (\mathbf{x} - \mu)' \Sigma^{-1} (\mathbf{x} - \mu).$$

For simplicity $d$ will be sometimes referred to as “distance”, although it should be kept in mind that it is actually a *squared* distance.
Then the multivariate analogue of $t_i^2$ where

$$t_i = \frac{x_i - \bar{x}}{s}$$

is the classical univariate outlyingness measure, is $D_i = d(x_i, \bar{x}, C)$ with $C = \text{Var}(X)$.

When $p = 1$ we have

$$D_i = t_i^2 \frac{n}{n - 1}.$$
It is known that if $\mathbf{x} \sim \mathcal{N}_p(\mu, \Sigma)$ then $d(\mathbf{x}, \mu, \Sigma) \sim \chi_p^2$.

Thus, assuming the estimates $\bar{x}$ and $C$ are close to their true values, we may examine the Q-Q plot of $D_i$ vs. the quantiles of a $\chi_p^2$ distribution and delete observations for which $D_i$ is “too high”.

This approach may be effective when there is a single outlier, but as in the case of location it can be useless when $n$ is small, and as in regression, several outliers may mask one another.
Example 2  This dataset contains, for each of 59 wines grown in the same region in Italy, the quantities of 13 constituents. The original purpose of the analysis was to classify wines from different cultivars by means of these measurements.

The upper row of the next Figure shows the plots of the classical squared distances as a function of observation number, and their Q-Q plot with respect to the $\chi^2_p$ distribution.

No clear outliers stand out.
Wine example: Mahalanobis distances vs. index number for classical and SR-05 estimates (left column), and Q-Q plots of distances (right column)
The lower row shows the results of using a robust estimate ("S estimate") to be defined later.

At least seven points stand out clearly:

The failure of the classical analysis in the upper row of the Figure shows that several outliers may "mask" one another.

These seven outliers have a strong influence on the results of the analysis.
Simple robust estimates of multivariate location can be obtained by applying a robust univariate location estimate to each coordinate, but this lacks affine equivariance.

For dispersion, there exist simple robust estimates of the covariance between two variables ("pairwise covariances") which could be used to construct a "robust covariance" matrix.

Apart from not being equivariant, the resulting matrix may not be positive semidefinite.

Non-equivariant procedures may also lack robustness when the data are very collinear.

In subsequent sections we will discuss a number of equivariant estimates that are robust analogues of the mean and covariance matrix.
They will be generally called location vectors and dispersion matrices.

The latter are also called robust covariance matrices in the literature.

Note that if the matrix $\Sigma$ with elements $\sigma_{jk}, \ j, k = 1, \ldots, p$ is a “robust covariance matrix”, then the matrix $R$ with elements

$$r_{jk} = \frac{\sigma_{jk}}{\sqrt{\sigma_{jj}\sigma_{kk}}}$$

is a robust analogue of the correlation matrix.
2 Breakdown and efficiency of multivariate estimates

2.1 Breakdown point

To define the breakdown point of \((\hat{\mu}, \hat{\Sigma})\) we must establish the meaning of "bounded away from the boundary of the parameter space".

For the location vector the parameter space is a finite-dimensional Euclidean space, and so "bounded away from the boundary" means simply that \(\hat{\mu}\) remains in a bounded set.
However the dispersion matrix has a more complex parameter space consisting of the set of symmetric nonnegative-definite matrices.

Each such matrix is characterized by the matrix of its eigenvectors and associated non-negative eigenvalues.

Thus “$\hat{\Sigma}$ bounded away from the boundary” is equivalent to the eigenvalues being bounded away from zero and infinity.

If we are interested in linear relationships among the variables, then it is not only dangerous that the largest eigenvalue becomes too large (“explosion”) but also that the smallest one becomes too small (“implosion”).

The first case is caused by outliers (observations far away from the bulk of the data),
the second by “inliers” (observations concentrated at some point or in general on a region of lower dimensionality).

For $0 \leq m \leq n$ call $\mathcal{Z}_m$ the set of “samples” $\mathcal{Z} = \{z_1, \ldots, z_n\}$ such that $\# \{z_i = x_i\} = m$, and call $\hat{\mu}(Z)$ and $\hat{\Sigma}(Z)$ the location vector and dispersion matrix estimates based on the sample $Z$.

The finite breakdown point of $(\hat{\mu}, \hat{\Sigma})$ is defined as $\varepsilon^* = m^*/n$ where $m^*$ is the largest $m$ such that there exist finite positive $a, b, c$ such that

$$\|\hat{\mu}(Z)\| \leq a \text{ and } b \leq \lambda_p(\hat{\Sigma}(Z)) \leq \lambda_1(\hat{\Sigma}(Z)) \leq c$$

for all $Z \in \mathcal{Z}_m$.

The asymptotic BP is defined analogously.
2.2 The multivariate exact fit property

A result analogous to that of the former Lecture holds for multivariate location and dispersion estimation.

Let the FBP of the affine equivariant estimate \((\hat{\mu}, \hat{\Sigma})\) be \(\varepsilon^* = m^*/n\).

Let the dataset contain \(q\) points on a hyperplane \(H = \{x : \beta'x = \gamma\}\) for some \(\beta \in \mathbb{R}^p\) and \(\gamma \in \mathbb{R}\).

If \(q \geq n - m^*\) then \(\hat{\mu} \in H\), and \(\hat{\Sigma}\beta = 0\).

That is, \(\hat{\mu}\) and \(\hat{\Sigma}\) fit the \(q\) points on \(H\) irrespective of the rest.
2.3 Efficiency

Call \((\hat{\mu}_n, \hat{\Sigma}_n)\) the estimates for a sample of size \(n\), and let \((\hat{\mu}_\infty, \hat{\Sigma}_\infty)\) be their asymptotic values.

All estimates considered in this chapter are consistent at the normal distribution in the following sense:

If \(x_i \sim N_p(\mu, \Sigma)\) then

\[
\hat{\mu}_\infty = \mu \quad \text{and} \quad \hat{\Sigma}_\infty = c\Sigma
\]

where \(c\) is a constant (if \(c = 1\) we have the usual definition of consistency).

Most estimates defined in this chapter are also asymptotically normal:

\[
\sqrt{n}(\hat{\mu}_n - \hat{\mu}_\infty) \xrightarrow{d} N_p(0, V_\mu), \quad \sqrt{n}\text{vec} \left(\hat{\Sigma}_n - \hat{\Sigma}_\infty\right) \xrightarrow{d} N_q(0, V_\Sigma),
\]
where \( q = p(p + 1)/2 \) and for a symmetric matrix \( \Sigma \), \( \text{vec}(\Sigma) \) is the vector containing the \( q \) elements of the upper triangle of \( \Sigma \).

The matrices \( V_\mu \) and \( V_\Sigma \) are the asymptotic covariance matrices of \( \hat{\mu} \) and \( \hat{\Sigma} \).

In general the estimate can be defined in such a way that \( c = 1 \) for a given model, e.g. the Multivariate Normal.

We consider the efficiency of \( \hat{\mu} \) when the data have a \( \mathcal{N}_p(\mu, \Sigma) \) distribution.

An affine equivariant location estimate \( \hat{\mu} \) has an asymptotic covariance matrix of the form

\[
V_\mu = v \Sigma,
\]

where \( v \) is a constant depending on the estimate.
In the case of the normal distribution MLE $\overline{x}$ we have $v = 1$ and the matrix $V_0$ is simply $\Sigma$, which results in $V_\mu^{-1}V_0 = v^{-1}I$ and $\text{eff}(\hat{\mu}) = 1/v$.

Thus the normal distribution efficiency of an affine equivariant location estimate is independent of $\mu$ and $\Sigma$.

The efficiency of $\hat{\Sigma}$ is much more complicated and will not be discussed here.
3 M estimates

Multivariate M-estimates will now be defined as for location, by generalizing maximum likelihood estimates.

Recall that in the univariate case it was possible to define separate robust equivariant estimates of location and of dispersion.

This is more complicated to do in the multivariate case, and if we want equivariant estimates it is better to estimate location and dispersion \textit{simultaneously}.

We shall develop the multivariate analogue of simultaneous M-estimates.
Recall that a multivariate normal density has the form

\[ f(x, \mu, \Sigma) = \frac{1}{\sqrt{|\Sigma|}} h(d(x, \mu, \Sigma)) \]

where

\[ h(s) = ce^{-s/2} \quad \text{with} \quad c = (2\pi)^{-p/2} \]

and

\[ d(x, \mu, \Sigma) = (x - \mu)'\Sigma^{-1}(x - \mu). \]

We note that the level sets of \( f \) are ellipsoidal surfaces.

In fact for any choice of positive \( h \) such that \( f \) integrates to one the level sets of \( f \) are ellipsoids, and so any density of this form is called *elliptically symmetric* (henceforth "elliptical" for short).
In the special case where $\mu = 0$ and $\Sigma = cI$ a density of this form is called *spherically symmetric* or *radial* (henceforth "spherical" for short).

It is easy to verify that the distribution $\mathcal{D}(x)$ is elliptical if and only if for some constant vector $a$ and matrix $A$, $\mathcal{D}(A(x - a))$ is spherical.

An important example of a non-normal elliptical distribution is the $p$-variate *Student distribution* with $\nu$ degrees of freedom ($0 < \nu < \infty$), which will be called $T_{p,\nu}$, and is obtained by the choice

$$h(s) = \frac{c}{(s + \nu)^{(p+\nu)/2}}$$

where $c$ is a constant.

The case $\nu = 1$ is called the multivariate Cauchy density, and the limiting case $\nu \to \infty$ yields the normal distribution.
If the mean (resp. the dispersion matrix) of an elliptical distribution exists, then it is equal to $\mu$ (resp. to a multiple of $\Sigma$).

Let $x_1, \ldots, x_n$ be an i.i.d. sample from an elliptical $f$. in which $h$ is assumed everywhere positive.

We shall calculate the MLE of $\mu$ and $\Sigma$. Put

$$\rho(s) = -2 \ln h(s) \text{ and } d_i = d(x_i, \hat{\mu}, \hat{\Sigma}).$$

Then maximizing the likelihood function yields the system of estimating equations.
\[
\sum_{i=1}^{n} W(d_i)(x_i - \hat{\mu}) = 0
\]

\[
\frac{1}{n} \sum_{i=1}^{n} W(d_i)(x_i - \hat{\mu})(x_i - \hat{\mu})' = \hat{\Sigma}
\]

with \( W = \rho' \)

For the normal distribution we have \( W \equiv 1 \) which yields the sample mean and sample covariance matrix for \( \hat{\mu} \) and \( \hat{\Sigma} \).

For the multivariate Student distribution we have

\[
W(d) = \frac{p + \nu}{d + \nu}.
\]

In general, we define M-estimates as solutions of
\[
\sum_{i=1}^{n} W_1(d_i)(x_i - \hat{\mu}) = 0
\]

\[
\frac{1}{n} \sum_{i=1}^{n} W_2(d_i)(x_i - \hat{\mu})(x_i - \hat{\mu})' = \hat{\Sigma}
\]

where the functions \( W_1 \) and \( W_2 \) need not be equal.

Note that we may interpret \( \hat{\Sigma} \) as a weighted covariance matrix we can express \( \hat{\mu} \) as the weighted mean

\[
\hat{\mu} = \frac{\sum_{i=1}^{n} W_1(d_i)x_i}{\sum_{i=1}^{n} W_1(d_i)}
\]

with weights depending on an outlyingness measure \( d_i \).
This is similar to the location case, in that we can express $\hat{\mu}$ as a weighted mean with data-dependent weights.

Existence and uniqueness of solutions were treated by Maronna (1976) and more generally by Tatsuoka and Tyler (2000).

Uniqueness of solutions of the estimating equations requires that $dW_2(d)$ be a nondecreasing function of $d$.

We shall call an M-estimate of location and dispersion \textit{monotone} if $dW_2(d)$ is nondecreasing, and \textit{redescending} otherwise.

Monotone M-estimates are defined as solutions to the estimating equations, while redescending ones must be defined by the minimization of some objective function, as happens with S-estimates.
For practical purposes monotone estimates are essentially unique, in the sense that all solutions to the M-estimating equations are consistent estimates.

If the $x_i$ are i.i.d. with distribution $F$ then under general assumptions when $n \to \infty$, monotone M-estimates defined converge in probability to the solution $(\hat{\mu}_\infty, \hat{\Sigma}_\infty)$ of

$$E_W^1(d)(x - \hat{\mu}_\infty) = 0,$$

$$E_W^2(d)(x - \hat{\mu}_\infty)(x - \hat{\mu}_\infty)' = \hat{\Sigma}_\infty$$

where $d = d(x, \hat{\mu}_\infty, \hat{\Sigma}_\infty)$.

Besides $\sqrt{n} \left( \hat{\mu} - \mu_\infty, \hat{\Sigma} - \Sigma_\infty \right)$ tends to a multivariate normal distribution, M estimates are affine equivariant, and therefore if $x$ has an elliptical distribution the asymptotic covariance matrix of $\hat{\mu}$ is proportional to $\Sigma_\infty$. 
3.1 Size and shape

If one dispersion matrix is a scalar multiple of another, i.e.

\[ \Sigma_2 = k \Sigma_1, \]

we say that they have the same shape, but different sizes.

Several important features of the distribution, such as correlations, principal components, and linear discriminant functions, depend only on shape.

Let \( \hat{\mu}_\infty \) and \( \hat{\Sigma}_\infty \) be the asymptotic values of location and dispersion estimates at an elliptical distribution \( F \).

In this case \( \hat{\mu}_\infty \) is equal to the center of symmetry \( \mu \), and \( \hat{\Sigma}_\infty \) is a constant multiple of \( \Sigma \),

\[ \hat{\Sigma}_\infty = c \Sigma, \]
with the proportionality constant $c$ depending on $F$ and on the estimator.

Most estimates to be considered are consistent for shape but not necessarily for size.

Therefore we must “scale” them so that size is also consistent.

This situation is similar to the scaling problem in with the MAD.

Consider estimate at the distribution $F = \mathbb{N}_p(\mu, \Sigma)$.

A simple approach to estimating the size of $\Sigma$ is based on noting that if

$$x \sim \mathbb{N}(\mu, \Sigma), \text{ then } d(x, \mu, \Sigma) \sim \chi^2_p,$$
and the fact that $\Sigma = c\hat{\Sigma}_\infty$ implies

$$cd(x, \mu, \Sigma) = d(x, \mu, \hat{\Sigma}_\infty).$$

Hence the empirical distribution of

$$\left\{d(x_1, \hat{\mu}, \hat{\Sigma}), ..., d(x_n, \hat{\mu}, \hat{\Sigma})\right\}$$

will resemble that of $d(x, \hat{\mu}_\infty, \hat{\Sigma}_\infty)$ which is $c\chi^2_p$

and so we may estimate $c$ robustly with

$$\hat{c} = \frac{\text{Med} \left\{d(x_1, \hat{\mu}, \hat{\Sigma}), ..., d(x_n, \hat{\mu}, \hat{\Sigma})\right\}}{\chi^2_p(0.5)}$$

where $\chi^2_p(\alpha)$ denotes the $\alpha$-quantile of the $\chi^2_p$ distribution.
3.2 Breakdown point

It is intuitively clear that robustness of the estimates requires that no term dominates the sums in the estimating equations, and to achieve this we assume

\[ W_1(d)\sqrt{d} \text{ and } W_2(d)d \text{ are bounded for } d \geq 0. \]

Let

\[ K = \sup_{d} W_2(d)d. \]

We first consider the asymptotic breakdown point, which is easier to deal with.

The “weak part” of joint M-estimates of \( \mu \) and \( \Sigma \) is the estimate \( \hat{\Sigma} \).
For if we take $\Sigma$ as known, then it is not difficult to prove that the asymptotic BP of $\hat{\mu}$ is $1/2$.

On the other hand in the case where $\mu$ is known the following result was obtained by Maronna (1976):

If the underlying distribution $F_0$ attributes zero mass to any hyperplane, then the asymptotic BP of a monotone M-estimate of $\Sigma$ verifies

$$\varepsilon^* \leq \frac{1}{p + 1}.$$

The finite BP is similar but the details are more involved.

Define a sample to be in general position if no hyperplane contains more than $p$ points.
Then Davies (1987) showed that the maximum finite BP of any equivariant estimate for a sample in general position is $m_{\text{max}}^*/n$ with

$$m_{\text{max}}^* = \left\lfloor \frac{n - p}{2} \right\rfloor.$$

It is therefore natural to search for estimates whose BP is nearer to this maximum BP than that of monotone M-estimates.
4 Estimates based on a robust scale

Just as with the regression estimates based on robust scales, where we aimed at making the residuals “small”, we shall define multivariate estimates of location and dispersion that make the distances $d_i$ “small”.

To this end we look for $\hat{\mu}$ and $\hat{\Sigma}$ minimizing some measure of “largeness” of $d(x, \hat{\mu}, \hat{\Sigma})$.

This can be trivially attained by letting the smallest eigenvalue of $\hat{\Sigma}$ tend to 0.

To prevent this we impose the constraint $|\hat{\Sigma}| = 1$.

Call $\mathcal{S}_p$ the set of symmetric positive definite $p \times p$-matrices.
For a dataset $X$ call $d \left(X, \hat{\mu}, \hat{\Sigma}\right)$ the vector with elements $d(x_i, \hat{\mu}, \hat{\Sigma}), i = 1, \ldots, n$, and let $\hat{\sigma}$ be a robust scale estimate.

Then we define the estimates $\hat{\mu}$ and $\hat{\Sigma}$ by

$$\hat{\sigma} \left(d \left(X, \hat{\mu}, \hat{\Sigma}\right)\right) = \min \text{ with } \hat{\mu} \in \mathbb{R}^p, \hat{\Sigma} \in \mathcal{S}_p, \left|\hat{\Sigma}\right| = 1.$$ 

It is easy to show that this type of estimates are equivariant.

An equivalent formulation of the above goal is to minimize $\left|\hat{\Sigma}\right|$ subject to a bound on $\hat{\sigma}$. 
4.1 The MVE estimate

The simplest case is to mimic the approach that results in the regression LMS estimate.

When $\hat{\sigma}$ is the sample median, the resulting location and dispersion matrix estimate is called the *minimum volume ellipsoid* (MVE) estimate.

The name stems from the fact that among all ellipsoids $\{x : d(x, \mu, \Sigma) \leq 1\}$ containing at least half of the data points, the one given by the MVE estimate has minimum volume, i.e., the minimum $|\Sigma|$.

The consistency rate of the MVE is the same slow rate as the LMS, namely only $n^{1/3}$, and hence is very inefficient.
4.2 S-estimates

To overcome the inefficiency of the MVE we consider a more general class of estimates called S-estimates (Davies, 1987), taking for $\hat{\sigma}$ an M-scale estimate that satisfies

$$
\frac{1}{n} \sum_{i=1}^{n} \rho \left( \frac{d_i}{\hat{\sigma}} \right) = \delta
$$

where $\rho$ is a smooth bounded $\rho$-function.

An S-estimate $\left( \hat{\mu}, \hat{\Sigma} \right)$ is an M-estimate in the sense that it satisfies estimating equations of the form

$$
\sum_{i=1}^{n} W \left( \frac{d_i}{\hat{\sigma}} \right) (x_i - \hat{\mu}) = 0,
$$
\[
\frac{1}{n} \sum_{i=1}^{n} W \left( \frac{d_i}{\hat{\sigma}} \right) (x_i - \hat{\mu})(x_i - \hat{\mu})' = c \hat{\Sigma},
\]

where

\[ W = \rho' \text{ and } \hat{\sigma} = \hat{\sigma}(d_1, \ldots, d_n), \]

and \( c \) is a scalar such that \( \left| \hat{\Sigma} \right| = 1 \).

Note however that if \( \rho \) is bounded (as is the usual case), \( dW(d) \) cannot be monotone (actually for the estimates usually employed \( W(d) \) vanishes for large \( d \)).

Hence the estimate is not a monotone M-estimate, and therefore the estimating equations yield only local minima of \( \hat{\sigma} \).
The choice $\rho(d) = d$ yields the average of the $d_i$ as scale estimate, in which case $W \equiv 1$ and hence
\[
\hat{\mu} = \overline{x}, \quad \hat{\Sigma} = \frac{C}{|C|^{1/p}}
\]
where $C$ is the sample covariance matrix.

It can be shown that if $\rho$ is differentiable, then for $S$-estimates the distribution of $\sqrt{n} \left( \hat{\mu} - \mu_\infty, \hat{\Sigma} - \Sigma_\infty \right)$ tends to a multivariate normal.

It can be shown that $S$ estimates can attain the maximum finite BP.

Note that the MVE can be considered as an $S$ estimate with discontinuous $\rho :$
\[
\rho(d) = 1(d > 1).
\]
A frequently used option is the bisquare multivariate S-estimate with
\[ \rho(t) = \min \left\{ 1, 1 - (1 - t)^3 \right\}, \]
which has weight function
\[ W(t) = 3 (1 - t)^2 I(t \leq 1). \]

Another useful estimate for large \( p \) corresponds to the “Rocke-type” \( \rho \)-function, which is a smoothed version of the \( \rho \)-function for the MVE. Its weight function is not monotonic. Details are omitted.

### 4.3 The MCD estimate

Another possibility is to use a trimmed scale for \( \hat{\sigma} \) instead of an M-scale, as was done to define the regression LTS estimate.
Let \( d(1) \leq \ldots \leq d(n) \) be the ordered values of the squared distances \( d_i = d(x_i, \mu, \Sigma) \).

Then for \( 1 < h < n \) define the trimmed scale of the squared distances as

\[
\hat{\sigma} = \sum_{i=1}^{h} d(i).
\]

An estimate \((\hat{\mu}, \hat{\Sigma})\) minimizing this trimmed scale is called a \textit{minimum covariance determinant} (MCD) estimate.

The reason for the name is the following: for each ellipsoid \( \{ x : d(x, t, V) \leq 1 \} \) containing at least \( h \) data points, compute the covariance matrix \( C \) of the data points in the ellipsoid.
If \((\hat{\mu}, \hat{\Sigma})\) is an MCD estimate, then the ellipsoid with \(t = \hat{\mu}\) and \(V\) equal to a scalar multiple of \(\hat{\Sigma}\) minimizes \(|C|\).

The MCD estimate can also attain the maximum BP.

5 The Stahel-Donoho estimate

Recall that the simplest approach to the detection of outliers in a univariate sample is the one given for location:

for each data point compute an “outlyingness measure” and identify those points having a “large” value of this measure.
The key idea for the extension of this approach to the multivariate case is that a multivariate outlier should be an outlier in some univariate projection.

More precisely, given a direction $a \in \mathbb{R}^p$ with $\|a\| = 1$, denote by

$$a'X = \{a'x_1, \ldots, a'x_n\}$$

the projection of the dataset $X$ along $a$.

Let $\hat{\mu}$ and $\hat{\sigma}$ be robust univariate location and dispersion statistics, e.g., the median and MAD respectively.

The outlyingness with respect to $X$ of a point $x \in \mathbb{R}^p$ along $a$ is defined by

$$t(x, a) = \frac{x' a - \hat{\mu}(a'X)}{\hat{\sigma}(a'X)}.$$
The outlyingness of \( x \) is then defined by

\[
t(x) = \max_a t(x, a).
\]

In the maximum above \( a \) ranges over the set \( \{||a|| = 1\} \), but in view of the equivariance of \( \hat{\mu} \) and \( \hat{\sigma} \), it is equivalent to take the set \( \{a \neq 0\} \).

The Stahel-Donoho estimate, proposed by Stahel (1981) and Donoho (1982), is a weighted mean and covariance matrix where the weight of \( x_i \) is a nonincreasing function of \( t(x_i) \).

More precisely, let \( W_1 \) and \( W_2 \) be two weight functions, and define

\[
\hat{\mu} = \frac{1}{\sum_{i=1}^{n} w_i 1} \sum_{i=1}^{n} w_i 1 x_i,
\]
\[ \hat{\Sigma} = \frac{1}{\sum_{i=1}^{n} w_{i2}} \sum_{i=1}^{n} w_{i2} (x_i - \hat{\mu})(x_i - \hat{\mu})' \]

with

\[ w_{ij} = W_j(t(x_i)), \quad j = 1, 2. \]

If \( y_i = Ax_i + b \), then it is easy to show that \( t(y_i) = t(x_i) \) (\( t \) is invariant) and hence the estimates are equivariant.

In order that no term dominates in the defining sums, it is clear that the weight functions must satisfy the conditions

\[ tW_1(t) \quad \text{and} \quad t^2W_2(t) \] are bounded for \( t \geq 0 \).

It can be shown that then the asymptotic BP is \( 1/2 \).
The choice of the weight functions is important for combining robustness and efficiency (no time for details).
6 Asymptotic bias

We now deal with data from a contaminated distribution $F = (1 - \varepsilon)F_0 + \varepsilon G$, where $F_0$ describes the “typical” data.

In order to define bias, we have to define which are the “true” parameters to estimate.

To fix ideas assume $F_0 = \mathcal{N}_p(\mu_0, \Sigma_0)$, but note that the following discussion applies to any other elliptical distribution.

Let $\hat{\mu}_\infty$ and $\hat{\Sigma}_\infty$ be the asymptotic values of location and dispersion estimates.

Defining a single measure of bias for a multidimensional estimate is more complicated than for regression.
Assume first that $\Sigma_0 = I$.

In this case the symmetry of the situation makes it natural to choose the Euclidean norm $\|\mu_\infty - \mu_0\|$ as a reasonable bias measure for location.

For the dispersion matrix size is relatively easy to adjust, and it will be most useful to focus on shape.

Thus we want to measure the discrepancy between $\hat{\Sigma}_\infty$ and scalar multiples of $I$.

The simplest way to do so is with the condition number, which is defined as the ratio of the largest to the smallest eigenvalue,

$$\text{cond}(\hat{\Sigma}_\infty) = \frac{\lambda_1(\hat{\Sigma}_\infty)}{\lambda_p(\hat{\Sigma}_\infty)}.$$
The condition number equals one if and only if $\hat{\Sigma}_\infty = cI$ for some $c \in R$.

Other functions of the eigenvalues may be used for measuring shape discrepancies.

In the special case of a spherical distribution the asymptotic value of an equivariant $\hat{\Sigma}$ is a scalar multiple of $I$, and so in this case there is no shape discrepancy.

For the case of an equivariant estimate and a general $\Sigma_0$ we want to define bias so that it is invariant under affine transformations, i.e., the bias does not change if $x$ is replaced by $Ax + b$.

To this end we define

\[
\text{bias}(\hat{\mu}) = \sqrt{(\hat{\mu}_\infty - \mu_0)' \Sigma_0^{-1} (\hat{\mu}_\infty - \mu_0)},
\]

\[
\text{bias}(\hat{\Sigma}) = \text{cond}(\Sigma_0^{-1} \hat{\Sigma}_\infty).
\]
It is easy to show that if the estimates are equivariant then $\text{bias}(\hat{\mu})$ does not depend upon either $\mu_0$ or $\Sigma_0$.

Hence to evaluate equivariant estimates we may without loss of generality take $\mu_0 = 0$ and $\Sigma_0 = I$.

The Table gives the maximum asymptotic biases of several multivariate location estimates at the multivariate normal distribution, for $\varepsilon = 0.1$ and $p = 5, 10$ and 20.

“S-D9” is the Stahel-Donoho estimate using "Huber" weights;

“Bisq” is the S-estimate with bisquare $\rho$ and BP=0.5.

“SR05” is an S estimate with "Rocke’s $\rho$-function".
Table 1: Maximum biases of multivariate estimates for contamination rate 0.10

<table>
<thead>
<tr>
<th></th>
<th>5</th>
<th>10</th>
<th>20</th>
</tr>
</thead>
<tbody>
<tr>
<td>MVE</td>
<td>0.73</td>
<td>0.75</td>
<td>0.77</td>
</tr>
<tr>
<td>Bisq</td>
<td>0.46</td>
<td>1.40</td>
<td>6.90</td>
</tr>
<tr>
<td>SR05</td>
<td>0.63</td>
<td>0.92</td>
<td>1.24</td>
</tr>
<tr>
<td>S-D9</td>
<td>0.52</td>
<td>1.07</td>
<td>2.47</td>
</tr>
<tr>
<td>MCD</td>
<td>0.94</td>
<td>1.97</td>
<td>7.00</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>Disp</td>
<td>6.9</td>
<td>9.4</td>
<td>15.0</td>
</tr>
<tr>
<td></td>
<td>4.05</td>
<td>19.31</td>
<td>357.42</td>
</tr>
<tr>
<td></td>
<td>4.77</td>
<td>9.52</td>
<td>23.90</td>
</tr>
</tbody>
</table>

The results show that the SR05 is the best competitors relative to the MVE, with somewhat smaller maximum biases than the MVE estimate for \( p = 5 \), and somewhat higher relative biases for \( p = 10 \) and \( 20 \).
7 Numerical computing of multivariate estimates

7.1 Monotone M-estimates

The estimating equations

\[
\sum_{i=1}^{n} W_1(d_i)(x_i - \hat{\mu}) = 0
\]

\[
\frac{1}{n} \sum_{i=1}^{n} W_2(d_i)(x_i - \hat{\mu})(x_i - \hat{\mu})' = \hat{\Sigma}
\]

yield an iterative algorithm similar to the one for regression.

Start with initial estimates $\hat{\mu}_0$ and $\hat{\Sigma}_0$, for example:
\( \hat{\mu}_0 \) = vector of coordinate-wise medians and
\( \hat{\Sigma}_0 \) = diagonal matrix with the squared normalized MADs of the variables in the diagonal.

At iteration \( k \) let \( d_{ki} = d(x_i, \hat{\mu}_k, \hat{\Sigma}_k) \) and compute
\[
\hat{\mu}_{k+1} = \frac{\sum_{i=1}^{n} W_1(d_{ki})x_i}{\sum_{i=1}^{n} W_1(d_{ki})}, \quad \hat{\Sigma}_{k+1} = \frac{1}{n} \sum_{i=1}^{n} W_2(d_{ki})(x_i - \hat{\mu}_{k+1})(x_i - \hat{\mu}_{k+1})'.
\]

The procedure is convergent.

Since the solution is unique for monotone M-estimates, the starting values influence only the number of iterations but not the end result.
7.2 Local solutions for S-estimates

Since local minima of $\hat{\sigma}$ are solutions of the M-estimating equations, a natural procedure to minimize $\hat{\sigma}$ is to use the former iterative procedure to solve the equations, with $W_1 = W_2$ equal to $W = \rho'$. 

It must be recalled that since $tW(t)$ is redescending, this pair of equations yields only a local minimum of $\sigma$, and hence the starting values are essential.

If the weight function $W$ is nonincreasing, then $\hat{\sigma}_k$ decreases at each step.

If $W$ is not monotonic (like with Rocke’s function), the iteration steps are not guaranteed to cause a decrease in $\hat{\sigma}_k$ at each step. 

However, the algorithm can be modified to insure a decrease at each iteration. The details are involved.
7.3 Subsampling for estimates based on a robust scale

The obvious procedure to generate an initial approximation to an estimate defined by the minimization of a scale similar to the general approach used for regression, in which the minimization problem is replaced by a finite one where the candidate estimates are sample means and covariance matrices of subsamples.

To obtain a finite set of candidate solutions take a subsample of size $p + 1$: $\{x_i : i \in J\}$ where the set $J \subset \{1, ..., n\}$ has $p + 1$ elements and compute
\[ \hat{\mu}_J = \text{ave}_{i \in J}(x_i) \text{ and } \hat{\Sigma}_J = \frac{\hat{C}_J}{|\hat{C}_J|^{1/p}}, \]

where \( \hat{C}_J \) is the covariance matrix of the subsample; and let

\[ d_J = \{d_{J_i} : i = 1, \ldots, n\}, \quad d_{J_i} = d(x_i, \hat{\mu}_J, \hat{\Sigma}_J). \]

Then the problem of minimizing \( \hat{\sigma} \) is replaced by the finite problem of minimizing \( \hat{\sigma}(d_J) \) over \( J \).

Since choosing all \( \binom{n}{p+1} \) subsamples is prohibitive unless both \( n \) and \( p \) are rather small, we choose \( N \) of them at random: \( \{J_k : k = 1, \ldots, N\} \).
Then the estimates are $\hat{\mu}_{J_k}, \hat{\Sigma}_{J_k}$ with

$$k^* = \arg \min_{k=1,\ldots,N} \hat{\sigma}(d_{J_k}).$$

If the sample contains a proportion $\varepsilon$ of outliers, the probability of least one “good” subsample is $1 - (1 - \alpha)^N$ where $\alpha = (1 - \varepsilon)^{p+1}$.

If we want this probability to be larger than $1 - \delta$ we must have

$$N \geq \frac{|\ln \delta|}{|\ln(1 - \alpha)|} \approx \frac{|\ln \delta|}{(1 - \varepsilon)^{p+1}}.$$

We have seen in the former Lecture the values of $N$ required as a function of $p$ and $\varepsilon$.  

7.4 The MVE

We have seen that the objective function of S-estimates can be decreased by iteration steps, and the same happens with the MCD.

However, no such improvements are known for the MVE, which makes the outcome of the subsampling procedure the only available approximation to the estimate.

In our book we propose an improved subsampling method (details are omitted).
7.5 Computation of S-estimates

Once we have initial values \( \hat{\mu}_0 \) and \( \hat{\Sigma}_0 \), an S-estimate is computed by means of the iterative procedures already described.

We present two approaches to compute \( \hat{\mu}_0 \) and \( \hat{\Sigma}_0 \).

The simplest approach is to obtain initial values of \( \mu_0 \), \( \Sigma_0 \) through subsampling and then apply the iterative algorithm.

Our preferred approach, however, proceeds as was done to compute the MM-estimates of regression; namely, start the iterative algorithm from a bias-robust but possibly inefficient estimate, which is computed through subsampling.
Since the last Table provides evidence that the MVE estimate has the smallest maximum bias, it is natural to think of using it as an initial estimate.

It is important to note that although the MVE estimate has the unattractive feature of a slow $n^{-1/3}$ rate of consistency.

This feature does not affect the efficiency of the local minimum which is the outcome of the iterative algorithm, since it satisfies the M-estimating equations.

If the equations for the asymptotic values $\hat{\mu}_\infty$, $\hat{\Sigma}_\infty$ have a unique solution, then all solutions converge to $\left(\hat{\mu}_\infty, \hat{\Sigma}_\infty\right)$ with order $n^{-1/2}$ rate.

Therefore we recommend use of the MVE as an initial estimate based on a subsampling approach, but using the improved method.
Simulation results show that this greatly improves the behavior of the estimates as compared to the simpler subsampling.

7.6 The MCD

Rousseeuw and van Driessen (1999) found a clever iterative algorithm for the MCD.

Details are omitted.
7.7 The Stahel-Donoho estimate

No exact algorithm for the Stahel-Donoho estimate is known.

To approximate the estimate we need a large number of directions, and these can be obtained by subsampling.

For each subsample \( J = \{x_{i_1}, \ldots, x_{i_p}\} \) of size \( p \), let \( a_J \) be a vector of norm one orthogonal to the hyperplane spanned by the subsample.

Then we generate \( N \) subsamples \( J_1, \ldots, J_N \) and replace the original outlyingness measure by

\[
\hat{t}(x) = \max_{k} t(x, a_{J_k})
\]

It is easy to show that \( \hat{t} \) is invariant under affine transformations, and hence the approximate estimate is equivariant.
8 Comparing estimates

Until now we have relied on asymptotic results, and we need some sense of the behavior of the estimates defined in the previous sections in finite sample sizes.

Since the behavior of robust estimates for finite $n$ is in general analytically intractable we must resort to simulation.

Recall that the performance of an estimate is a combination of its bias and variability.

The relative performances of two estimates depend on the underlying distribution and also on the sample size.
Since the variability tends to zero when $n \to \infty$ while the bias remains essentially constant, an estimate with low bias but high variability may be very good for very large $n$ but bad for moderate $n$.

Hence no estimate can be “best” under all circumstances.

Computing time is also an element to be taken into account.

The simulations presented in this section are all based on an underlying a $p$-variate normal distribution $N_p(0, I)$ with 10% point-mass contamination located at distance $k$ of the origin, with $k$ ranging on a suitable interval.

The criterion we use to compare the estimates is the root mean square error (RMSE) of the location estimate.
The relative performances of the dispersion matrices are similar to those of the location vectors, and are hence not shown here.

We first compare the two versions of MVE mentioned in a former Section:

- simple subsampling (labeled “MVE-1”), and

- subsampling with the improvement (labeled “MVE-2”).

The next Figure displays the results for $p = 10$ with 500 subsamples, and the advantages of MVE-2 are clear. Hence in the remainder of the discussion only the improved version MVE-2 will be employed.
RMSE of location estimate from MVE with simple (MVE-1) and improved subsampling (MVE-2), for $p = 10$.

Now we compare two ways of approximating the bisquare S-estimate as described in a former Section:
• starting from the older "naive" subsampling based estimate (labeled "Bisqu-1"), and

• starting from the improved MVE-2 estimate (labeled "Bisqu-2").

The next Figure shows for $p = 10$ the advantages of a good starting point.

All S-estimates considered henceforth are computed with the MVE-2 as starting point.
RMSE of bisquare location estimate from pure subsampling (Bisq-1) and starting from fast MVE (Bisq-2) for $p = 10$. 
To compare the behaviors of several estimates we performed a simulation experiment, which included

- The classical sample mean and covariance matrix
- The MVE-2 and MCD estimates
- The S-estimate with bisquare function
- The S-estimate with Rocke-type $\rho$-function SR-05
The sampling situations were $\varepsilon$-contaminated $p$-variate normal distributions $N_p(0, I)$, with $\varepsilon = 0.10$, $p = 5$, and 20, and with $n = 10p$.

In each plot “$k = 0$” corresponds to the case $\varepsilon = 0$ so that we can compare normal distribution efficiencies.

The next Figure shows the results for $p = 5$ and $n = 50$. 
Simulation RMSEs of location estimates for $p = 5$ and $n = 50$

The bisquare estimate is clearly more robust and more efficient than the others.

The next Figure shows the results for $p = 20$ and $n = 200$. 
Simulation RMSEs for multivariate location estimates with $p = 20$ and $n = 200$
Here SR-05 is clearly better than its competitors. The values for the MCD eventually drop for large $k$ (the maximum asymptotic bias of MCD for $p = 20$ is attained for $k = 17$).

As a consequence of these simulations and of the former asymptotic results we propose the use of S-estimates, since they give an adequate balance between bias and variability and can be computed in feasible times.

As a simple rule, we recommend the bisquare for $p < 10$ and SR-05 for $p \geq 10$. 
9 Faster estimates based on pairwise robust covariances

Estimates based on a subsampling approach will be too slow when \( p \) is large, for example when \( p \) is on the order of a few hundred.

Much faster estimators can be obtained if equivariance is given up.

The simplest approaches for location and dispersion are respectively to apply a robust location estimate to each coordinate and a robust estimate of covariance to each pair of variables.

Such pairwise robust covariance estimates are easy to compute, but unfortunately the resulting dispersion matrix lacks affine equivariance and positive definiteness.
Besides, such estimates for location and dispersion may lack both bias robustness and high normal efficiency if the data are very correlated.

This is because the coordinatewise location estimates need to incorporate the correlation structure for full efficiency at the normal distribution, and because the pairwise covariance estimates may fail to downweight higher-dimensional outliers.

A simple way to define a robust covariance between two random variables $x, y$ is by truncation or rejection.

Let $\psi$ be a bounded monotone or redescending $\psi$-function, and $\mu(.)$ and $\sigma(.)$ robust location and dispersion statistics.
Then robust correlations and covariances can be defined as

$$\text{RCov}(x, y) = \sigma(x)\sigma(y)\mathbb{E}\left[\psi\left(\frac{x - \mu(x)}{\sigma(x)}\right)\psi\left(\frac{y - \mu(y)}{\sigma(y)}\right)\right],$$

$$\text{RCorr}(x, y) = \frac{\text{RCov}(x, y)}{[\text{RCov}(x, x)\text{RCov}(y, y)]^{1/2}}.$$ 

This definition satisfies $\text{RCorr}(x, x) = 1$.

When $\psi(x) = \text{sgn}(x)$ and $\mu$ is the median, $\text{RCov}$ and $\text{RCorr}$ are called the quadrant correlation and covariance estimates.

The sample versions thereof are obtained by replacing the expectation by the average and $\mu$ and $\sigma$ by their estimates $\hat{\mu}$ and $\hat{\sigma}$. 
These estimates are not consistent under a given model.

In particular, if $D(x, y)$ is bivariate normal with correlation $\rho$ and $\psi$ is monotone, then the value $\rho_R$ of $RCorr(x, y)$ is an increasing function $\rho_R = g(\rho)$ of $\rho$ which can be computed.

Then, the estimate $\hat{\rho}_R$ of $\rho_R$ can be corrected to ensure consistency at the normal model by using the inverse transformation $\hat{\rho} = g^{-1}(\hat{\rho}_R)$.

Another robust pairwise covariance initially proposed by Gnanadesikan and Kettenring (1972) is based on the identity

$$
\text{Cov}(x, y) = \frac{1}{4} \left( \text{SD} (x + y)^2 - \text{SD} (x - y)^2 \right).
$$
They proposed to define a robust correlation by replacing the standard deviation by a robust dispersion $\sigma$ (they chose a trimmed standard deviation):

$$\text{RCorr}(x, y) = \frac{1}{4} \left( \sigma \left( \frac{x}{\sigma(x)} + \frac{y}{\sigma(y)} \right)^2 - \sigma \left( \frac{x}{\sigma(x)} - \frac{y}{\sigma(y)} \right)^2 \right)$$

and a robust covariance is defined by

$$\text{RCov}(x, y) = \sigma(x)\sigma(y) \text{RCorr}(x, y).$$

The latter satisfies

$$\text{RCov}(t_1 x, t_2 y) = t_1 t_2 \text{RCov}(x, y) \text{ for all } t_1, t_2 \in R$$

and

$$\text{RCov}(x, x) = \sigma(x)^2.$$

The above pairwise robust covariances can be used in the obvious way to define a “robust correlation (or covariance) matrix” of a random vector $x = (x_1, \ldots, x_p)'$. 
The resulting dispersion matrix is symmetric but not necessarily positive semi-definite, and is not affine equivariant.

Maronna and Zamar (2002) show that a simple modification of Gnanadesikan and Kettenring’s approach yields a positive definite matrix and “approximately equivariant” estimates of location and dispersion. Details are omitted.
Principal components analysis (PCA) is a widely used method for dimensionality reduction.

Let $x$ be a $p$-dimensional random vector with mean $\mu$ and covariance matrix $\Sigma$.

The first principal component is the univariate projection of maximum variance.

More precisely, it is the linear combination $x'b_1$ where $b_1$ (called the first principal direction) is the vector $b$ such that

$$\text{Var}(b'x) = \max \text{ subject to } \|b\| = 1.$$
The second principal component is $x'b_2$ where $b_2$ (the second principal direction) satisfies the above condition with $b_2'b_1 = 0$, and so on.

Call $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$ the eigenvalues of $\Sigma$.

Then $b_1, \ldots, b_p$ are the respective eigenvectors and $\text{Var}(b_j'x) = \lambda_j$.

The number $q$ of components can be chosen on the basis of the “proportion of unexplained variance”

$$\frac{\sum_{j=q+1}^{p} \lambda_j}{\sum_{j=1}^{p} \lambda_j}.$$

PCA can be viewed in an alternative geometric form in the spirit of regression modeling.
Consider finding a $q$-dimensional hyperplane $H$ such the orthogonal distance of $x$ to $H$ is “smallest”, in the following sense.

Call $\hat{x}_H$ the point of $H$ closest in Euclidean distance to $x$, i.e., such that

$$\hat{x}_H = \arg \min_{z \in H} \|x - z\| .$$

Then we look for $H^*$ such that

$$E \|x - \hat{x}_{H^*}\|^2 = \min .$$

It can be shown that $H^*$ contains the mean $\mu$ and has the directions of the first $q$ eigenvectors $b_1, \ldots, b_q$.

Note that the results of PCA are not invariant under general affine transformations, in particular under changes in the units of the variables.
Doing PCA implies that we consider the Euclidean norm to be a sensible measure of distance, and this may require a previous rescaling of the variables.

PCA is however invariant under orthogonal transformations, i.e., transformations that do not change Euclidean distances.

Given a dataset $X = \{x_1, ..., x_n\}$, the sample principal components are computed by replacing $\mu$ and $\Sigma$ by the sample mean and covariance matrix.
For each observation \( x_i \) we compute the scores \( z_{ij} = (x_i - \bar{x})'b_j \) and the distances

\[
d_{\hat{H},i} = \sum_{j=q+1}^{p} z_{ij}^2 = \left\| x_i - \hat{x}_{\hat{H},i} \right\|^2, \quad d_{C,i} = \sum_{j=1}^{q} z_{ij}^2,
\]

where \( \hat{H} \) is the estimated hyperplane.

A simple data analytic tool similar to the plot of residuals versus fitted values in regression, is to plot \( d_{\hat{H},i} \) vs. \( d_{C,i} \).

As can be expected, outliers may have a distorting effect on the results.

For instance consider the initial Example:
Biochemical data

Here the first principal component of the correlation matrix of the data explains 75% of the variability, while after deleting the atypical point it explains 90%.
The simplest way to deal with this problem is to replace $\bar{x}$ and $\text{Var}(X)$ with robust estimates $\hat{\mu}$ and $\hat{\Sigma}$ of multivariate location and dispersion.

Note that the results depend only on the shape of $\hat{\Sigma}$.

However, better results can be obtained by taking advantage of the particular features of PCA.

The next Section describes a simple and fast method that works well for large datasets.
10.1 Spherical principal components

In this section we describe a simple but effective approach proposed by Locantore, Marron, Simpson, Tripoli, Zhang, and Cohen (1999), called Spherical Principal Components (SPC).

Let $\mu$ be a robust multivariate location estimate, and compute

$$y_i = \begin{cases} 
  \frac{(x_i - \mu)}{||x_i - \mu||} & \text{if } x_i \neq \mu \\
  0 & \text{otherwise}
\end{cases}$$

(i.e., the $y_i$'s are the $x_i$'s “sphericized”).

Let $\hat{V}$ the sample covariance matrix of the $y_i$, with corresponding eigenvectors $b_j$ ($j = 1, \ldots, p$).
Now compute

$$\hat{\lambda}_j = \hat{\sigma}(x'b_j)^2$$

where $\hat{\sigma}$ is a robust dispersion estimate (such as the MAD).

Call $\hat{\lambda}_{(j)}$ the sorted $\lambda$’s: $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$, and $b_{(j)}$ the corresponding eigenvectors.

Then the first $q$ principal directions are given by the $b_{(j)}$, $j = 1, \ldots, q$, and the respective “robust proportion of unexplained variance” is given by

$$\frac{\sum_{j=q+1}^{p} \lambda_{(j)}}{\sum_{j=1}^{p} \lambda_{(j)}}.$$

Boente and Fraiman (1999) showed that if $x$ is elliptical, then the eigenvectors $t_1, \ldots, t_p$ (but not the eigenvalues!) of the covariance matrix of $y$ (i.e., its principal axes) coincide with those of its covariance matrix $\Sigma$. 
They showed furthermore that if $\sigma\left(\cdot\right)$ is any dispersion statistic then the values $\sigma\left(x't_j\right)^2$ are proportional to the eigenvalues of $\Sigma$.

Therefore SPC is consistent at least for elliptical data.

The simplest choice for $\hat{\mu}$ is the coordinatewise median.

An orthogonal equivariant choice for $\hat{\mu}$ is the “space median”:

$$\hat{\mu} = \arg \min_\mu \sum_{i=1}^n \|x_i - \mu\|.$$  

Note that this is an M-estimate, and can therefore be easily computed,

It follows that this estimate has BP $= 0.5$. 
This SPC procedure is deterministic and very fast, and it can be computed with collinear data without any special adjustments.

Despite its simplicity, simulations by Maronna (2005) show that this SPC method performs very well.

**Example 3** *This dataset corresponds to a study in automatic vehicle recognition. Each of the 218 rows corresponds to a view of a bus silhouette, and contains 18 attributes of the image. The SD’s are in general much larger than the respective MADN’s. The latter vary between 0 (for variable 9) to 34. Hence it was decided to exclude variable 9 and to normalize the data by dividing the remaining variables by their MADN’s.*
The Table shows the proportions of unexplained variability as a function of the number $q$ of components, for the classical PCA and for SPC.

It would seem that since the classical method has smaller unexplained variability than the robust method, classical PCA gives a better representation.

However, this is not the case.
The next Table gives the quantiles of the distances to the hyperplane $d_{H,i}$ for $q = 3$, and the Figure compares the logs of the respective ordered values (the log scale was used because of the extremely large outliers).
Bus image data: QQ plot of logs of distances to hyperplane \( (q = 3) \) from classical and robust estimates. The line is the identity diagonal.
It is seen in the Figure that the hyperplane from the robust fit has in general smaller distances to the data points, except for some clearly outlying ones.

On the other hand in the former Table the classical estimate seems to perform better than the robust one.

The reason is that the two estimates use different measures of variability.
The classical procedure uses variances which are influenced by the outliers, and so large outliers in the direction of the first principal axes will inflate the corresponding variances and hence increase their proportion of explained variability.

On the other hand the robust spherical principal components uses a robust dispersion measure which is free of this drawback, and gives a more accurate measure of the unexplained variability for the bulk of the data.